

IDR
Internet-Draft
Intended status: Informational
Expires: January 15, 2013

P. Lapukhov
Microsoft Corp.
A. Premji
Arista Networks
July 14, 2012

Using BGP for routing in large-scale data centers
draft-lapukhov-bgp-routing-large-dc-01

Abstract

Some service providers build and operate data centers that support over 100,000 servers. In this document, such data-centers are referred to as "large-scale" data centers to differentiate them the from more common smaller infrastructures. The data centers of this scale have a unique set of network requirements, with emphasis on operational simplicity and network stability.

This document attempts to summarize the authors' experiences in designing and supporting large data centers, using BGP as the only control-plane protocol. The intent here is to describe a proven and stable routing design that could be leveraged by others in the industry.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 15, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal

Provisions Relating to IETF Documents
(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Traditional data center designs	3
2.1. Layer 2 Designs	3
2.2. Fully routed network designs	4
3. Document structure	5
4. Network design requirements	5
4.1. Traffic patterns	5
4.2. CAPEX minimization	6
4.3. OPEX minimization	6
4.4. Traffic Engineering	7
5. Requirement List	7
6. Network topology	7
6.1. Clos topology overview	8
6.2. Clos topology properties	8
6.3. Scaling Clos topology	9
7. Routing design	10
7.1. Choosing the routing protocol	10
7.2. BGP configuration for Clos topology	11
7.2.1. BGP Autonomous System numbering layout	11
7.2.2. Non-unique private BGP ASN's	12
7.2.3. Prefix advertisement	13
7.2.4. External connectivity	13
7.3. ECMP Considerations	14
7.3.1. Basic ECMP	14
7.3.2. BGP ECMP over multiple ASN	15
7.4. BGP convergence properties	16
7.4.1. Convergence timing	16
7.4.2. Failure impact scope	16
7.4.3. Third-party route injection	17
8. Security Considerations	17
9. IANA Considerations	17
10. Acknowledgements	17
11. Informative References	18
Authors' Addresses	19

1. Introduction

This document presents a practical routing design that can be used in large-scale data centers. Such data centers, also known as hyper-scale or warehouse scale data centers, have a unique attribute of supporting over a 100,000 end hosts. In order to support networks of such scale, operators are revisiting networking designs and platforms to address this need.. Contrary to the more traditional data center designs, the approach presented in this document does not have any dependency on building a large Layer-2 domain and instead relies on routing at every layer in the network. Implementing a pure Layer-3 design using BGP further ensures broad vendor support and almost guarantees interoperability between vendors given that BGP is one of the most widely deployed protocols on the Internet.

2. Traditional data center designs

This section provides an overview of two types of traditional data center designs - Layer-2 and fully routed Layer-3 topologies.

2.1. Layer 2 Designs

In the networking industry, a common design choice for data centers is to use a mix of Ethernet-based Layer 2 technologies. Network topologies typically look like a tree with redundant uplinks and three levels of hierarchy commonly named Core , Aggregation and Access layers (see Figure 1). To accommodate bandwidth demands, every next level has higher port density and bandwidth capacity, moving upwards in the topology. To keep terminology uniform, in this document, these topology layers will be referred to as "tiers", e.g. Tier 1, Tier 2 and Tier 3 instead of Core, Aggregation or Access layers.

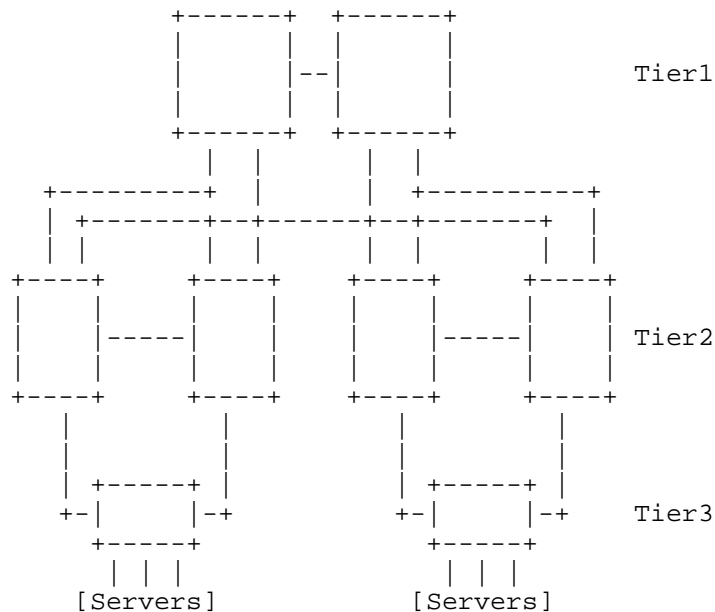


Figure 1: Typical Data Center network layout

IP routing is normally used only at the upper layers in the topology, e.g. Tier 1 or Tier 2. Some of the reasons for introducing such large (sometimes called stretched) layer-2 domains are:

- o Supporting legacy applications that may require direct Layer 2 adjacency or use non-IP protocols
- o Seamless mobility for virtual machines, to allow the preservation of IP addresses when a virtual machine moves across physical hosts
- o Simplified IP addressing - less IP subnets is required for the data-center
- o Application load-balancing may require direct layer-2 reachability to perform certain functions such as Level 2 Direct Server Return (DSR)

2.2. Fully routed network designs

Network designs that leverage IP routing down to the access layer (Tier 3) of the network have gained popularity as well. The main benefit of such designs is improved network stability and scalability, as a result of confining L2 broadcast domains. A common choice of routing protocol for data center designs would be an IGP, such as OSPF or ISIS. As data centers grow in scale, and server count exceeds tens of thousands, such fully routed designs become

more attractive.

Although BGP is the de-facto standard protocol for routing on the Internet, having wide support from both the vendor and service provider communities, it is not generally deployed in data centers for a number of reasons:

- o BGP is perceived as a "WAN only protocol only" and not often considered for enterprise or data center applications.
- o BGP is believed to have a "much slower" routing convergence than traditional IGPs.
- o BGP deployment within an Autonomous System (iBGP mesh) is assumed to have a dependency on the presence of an IGP, which assists with recursive next-hop resolution.
- o BGP is perceived to require significant configuration overhead and does not support any form of neighbor auto-discovery.

In this document we demonstrate a practical approach for using BGP as the single routing protocol for data center networks.

3. Document structure

The remaining of this document is organized as following. First the design requirements for large scale data centers are presented. Next, the document gives an overview of Clos network topology and its properties. After that, the reasons for selecting BGP as the single routing protocols are presented. Finally, the document discusses the design in more details and covers specific BGP policy features.

4. Network design requirements

This section describes and summarizes network design requirement for a large-scale data center.

4.1. Traffic patterns

The primary requirement when building an interconnection network for large number of servers is to accommodate application bandwidth and latency requirements. Until recently it was quite common to see traffic flows mostly entering and leaving the data center (also known as north-south traffic) There were no intense, highly meshed flows or traffic patterns between the machines within the same tier. As a result, traditional "tree" topologies were sufficient to accommodate such flows, even with high oversubscription ratios in network equipment. If more bandwidth was required, it was added by "scaling up" the network elements, by upgrading line-cards or switch fabrics.

In contrast, large-scale data centers often host applications that generate significant amount of server to server traffic, also known as "east-west" traffic. Examples of such applications could be compute clusters such as Hadoop or live virtual machine migrations. Scaling up traditional tree topologies to match these bandwidth demands becomes either too expensive or impossible due to physical limitations.

4.2. CAPEX minimization

The cost of the network infrastructure alone (CAPEX) constitutes about 10-15% of total data center expenditure [GREENBERG2009]. However, The absolute cost is significant, and there is a need to constantly drive down the cost of networking elements themselves. This can be accomplished in two ways:

- o Unifying all network elements, preferably using the same hardware type or even the same device. This allows for bulk purchases with discounted pricing.
- o Driving costs down by introducing multiple network equipment vendors.

In order to allow for vendor diversity, it is important to minimize the software feature requirements for the network elements. Furthermore, this strategy provides the maximum flexibility of vendor equipment choices while enforcing interoperability using open standards

4.3. OPEX minimization

Operating large scale infrastructure could be expensive, provide that larger amount of elements will statistically fail more often. Having a simpler design and operating using a limited software feature-set ensures that failures will mostly result from hardware malfunction and not software issues.

An important aspect of OPEX minimization is reducing size of failure domains in the network. Ethernet networks are known to be susceptible to broadcast or unicast storms. The use of a fully routed design significantly reduces the size of the data-plane failure domains (e.g. limits to Tier-3 switches only). However, such designs also introduce the problem of distributed control-plane failures. This calls for simpler control-plane protocols that are expected to have less chances of network meltdown.

4.4. Traffic Engineering

In any data center, application load-balancing is a critical function performed by network devices. Traditionally, load-balancers are deployed as dedicated devices in the traffic forwarding path. The problem arises in scaling load-balancers under growing traffic demand. A preferable solution would be able to scale load-balancing layer horizontally, by adding more of the uniform nodes and distributing incoming traffic across these nodes

In situation like this, an ideal choice would to use network infrastructure itself to distribute traffic across a group of load-balancers. A combination of features such as Anycast prefix advertisement [RFC4786] along with Equal Cost Multipath (ECMP) functionality could be used to accomplish this. To allow for more granular load-distribution, it is beneficial for the network to support the ability to perform controlled per-hop traffic engineering. For example, it is beneficial to directly control the ECMP next-hop set for anycast prefixes at every level of network hierarchy.

5. Requirement List

This section summarizes the list of requirements, based on the discussion so far:

- o REQ1: Select a network topology where capacity could be scaled "horizontally" by adding more links and network switches of the same type, without requiring an upgrade to the network elements themselves.
- o REQ2: Define a narrow set of software features/protocols supported by a multitude of networking equipment vendors.
- o REQ3: Among the network protocols, choose the one that has a simpler implementation in terms of minimal programming code complexity.
- o REQ4: The network routing protocol should allow for explicit control of the routing prefix next-hop set on per-hop basis.

6. Network topology

This section outlines the most common choice for horizontally scalable topology in large scale data centers.

6.1. Clos topology overview

A common choice for a horizontally scalable topology is a folded Clos topology, sometimes called "fat-tree" (see, for example, [INTERCON] and [ALFARES2008]). This topology features odd number of stages (dimensions) and is commonly made of the same uniform elements, e.g. switches with the same port count. Therefore, the choice of Clos topology satisfies both REQ1 and REQ2. See Figure 2 below for an example of folded 3-stage Clos topology:

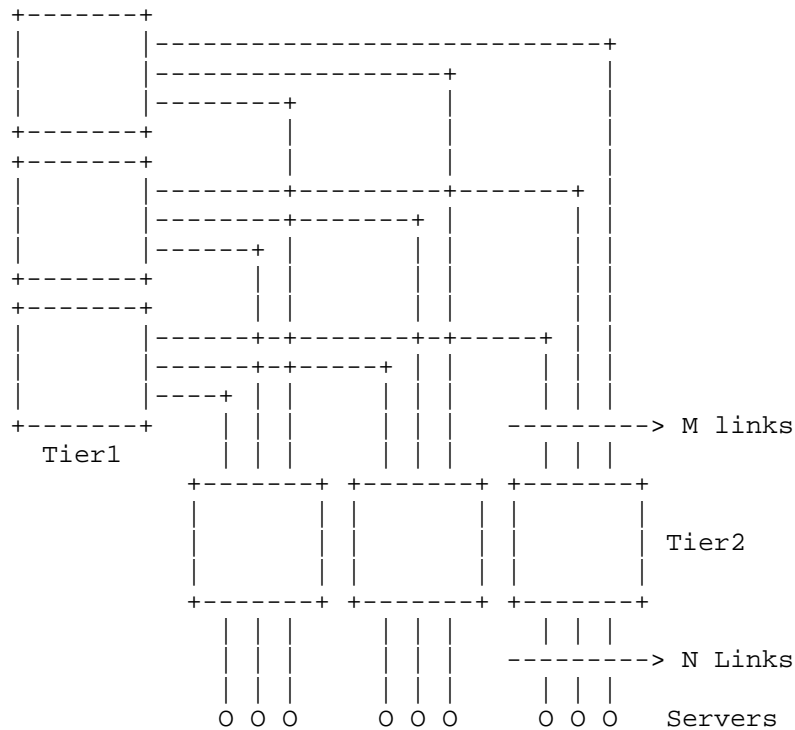


Figure 2: 3-Stage Folded Clos topology

In the networking industry, a topology like this is sometimes referred to as "Leaf and Spine" network, where "Spine" is the name given to the middle stage of the Clos topology (Tier 1) and "Leaf" is the name of input/output stage (Tier 2). However, for consistency, we will refer to these layers as "Tier n".

6.2. Clos topology properties

The following are some key properties of the Clos topology:

- o Topology is fully non-blocking (or more accurately - non-interfering) if $M \geq N$ and oversubscribed by a factor of N/M otherwise. Here M and N is the uplink and downlink port count respectively, for Tier 2 switch, as shown on Figure 2
- o Implementing Clos topology requires a routing protocol supporting ECMP with the fan-out of M or more
- o Every Tier 1 device has exactly one path to every end host (server) in this topology
- o Traffic flowing from server to server is naturally load-balanced over all available paths using simple ECMP behavior

6.3. Scaling Clos topology

A Clos topology could be scaled either by increasing network switch port count or adding more stages, e.g. moving to a 5-stage Clos, as illustrated on Figure 3 below:

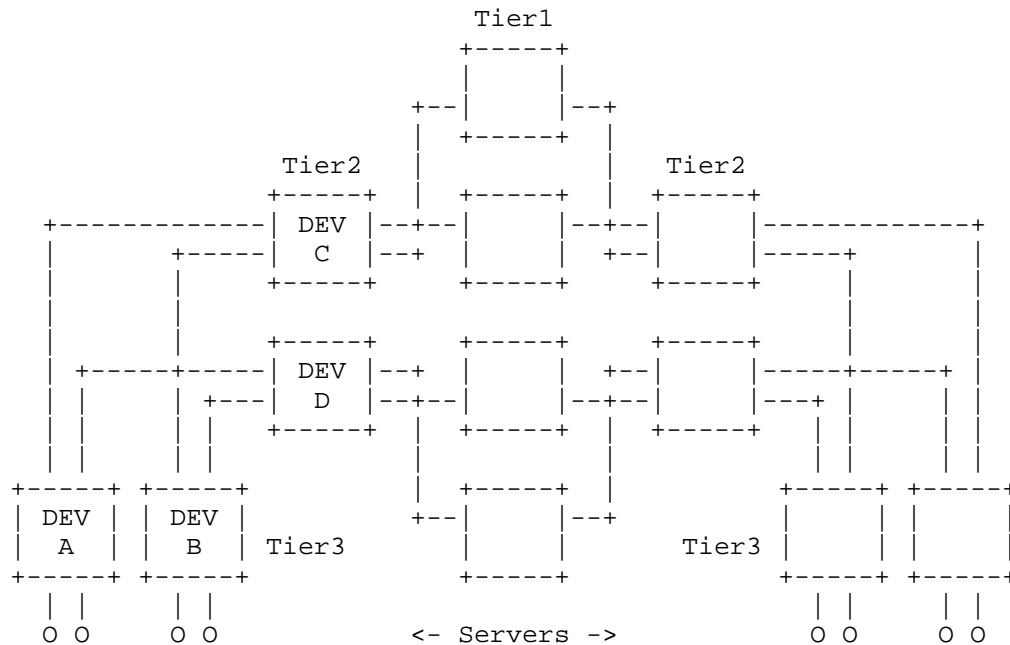


Figure 3: 5-Stage Clos topology

The topology on Figure 3 is built from switches with port count of 4 and provides full bisection bandwidth to all connected servers. We will refer to the collection of directly connected Tier 2 and Tier 3 switches as a "cluster" in this document. For example, devices A, B, C, and D on Figure 3 form a cluster.

In practice, the Tier 3 level of the network (typically top of rack switches, or ToRs) is where oversubscription is introduced to allow for packaging of more servers in data center. The main reason to limit oversubscription at a single layer of the network is to simplify application development that would otherwise need to account for two bandwidth pools: within the same access switch (e.g. rack) and outside of the local switch. Since oversubscription itself does not have any effect on routing, we will not be discussing it further in this document.

7. Routing design

This section discusses the motivation for choosing BGP as the routing protocol and BGP configuration for routing in Clos topology.

7.1. Choosing the routing protocol

The set of requirements discussed earlier call for a single routing protocol (REQ2) to reduce complexity and interdependencies. While it is common to rely on an IGP in this situation, the document proposes the use of BGP only. The advantages of using BGP are discussed below.

- o BGP inherently has less complexity within its protocol design - internal data structures and state-machines are simpler when compared to a link-state IGP. For example, instead of implementing adjacency formation, adjacency maintenance and/or flow-control, BGP simply relies on TCP as the underlying transport. This fulfills REQ1 and REQ2.
- o BGP information flooding overhead is less when compared to link-state IGPs. Indeed, since every BGP router normally re-calculates and propagates best-paths only, a network failure is masked as soon as the BGP speaker finds an alternate path. In contrary, the event propagation scope of a link-state IGP is single flooding domain, regardless of the failure type. Furthermore, all well-known link-state IGPs feature periodic refresh updates, while BGP does not expire routing state.
- o BGP supports third-party (recursively resolved) next-hops. This allows for ECMP or forwarding based on customer-defined forwarding paths. This satisfied REQ4 stated above. Some IGPs, such as OSPF, support similar functionality using special concepts such as "Forwarding Address", but do not satisfy other requirement, such as protocol simplicity.
- o Vanilla BGP configuration, without routing policies, is easier to troubleshoot for network reachability issues. For example, it is straightforward to dump contents of LocRIB and compare it to the router's RIB and FIB. Furthermore, every BGP neighbor has

corresponding AdjRIBIn and AdjRIBOut structures with incoming/outgoing NRLI information that could be easily correlated on both sides of the BGP peering session. Thus BGP fully satisfies REQ3.

7.2. BGP configuration for Clos topology

Topologies that have more than 5 stages are very uncommon due to the large numbers of interconnects required by such a design.

7.2.1. BGP Autonomous System numbering layout

The diagram below illustrates suggests BGP Autonomous System Number (BGP ASN) allocation scheme. The following is a list of guidelines that can be used:

- o All BGP peering sessions are external BGP (eBGP) established over direct point-to-point links interconnecting the network nodes.
- o 16-bit (two octet) BGP ASNs are used, since these are widely supported and have better vendor interoperability (e.g. no need to support BGP capability negotiation).
- o Private BGP ASNs from the range 64512-64534 are used so as to avoid ASN conflicts. The private ASN stripping feature can be leveraged as a result (see below).
- o A single BGP ASN is allocated to the Clos middle stage ("Tier 1"), e.g. ASN 64534 as shown in Figure 4
- o Unique BGP ASN is allocated per group of "Tier 2" switches. All Tier 2 switches in the same group share the BGP ASN.
- o Unique BGP ASN is allocated to every Tier 3 switch (e.g. ToR) in this topology.

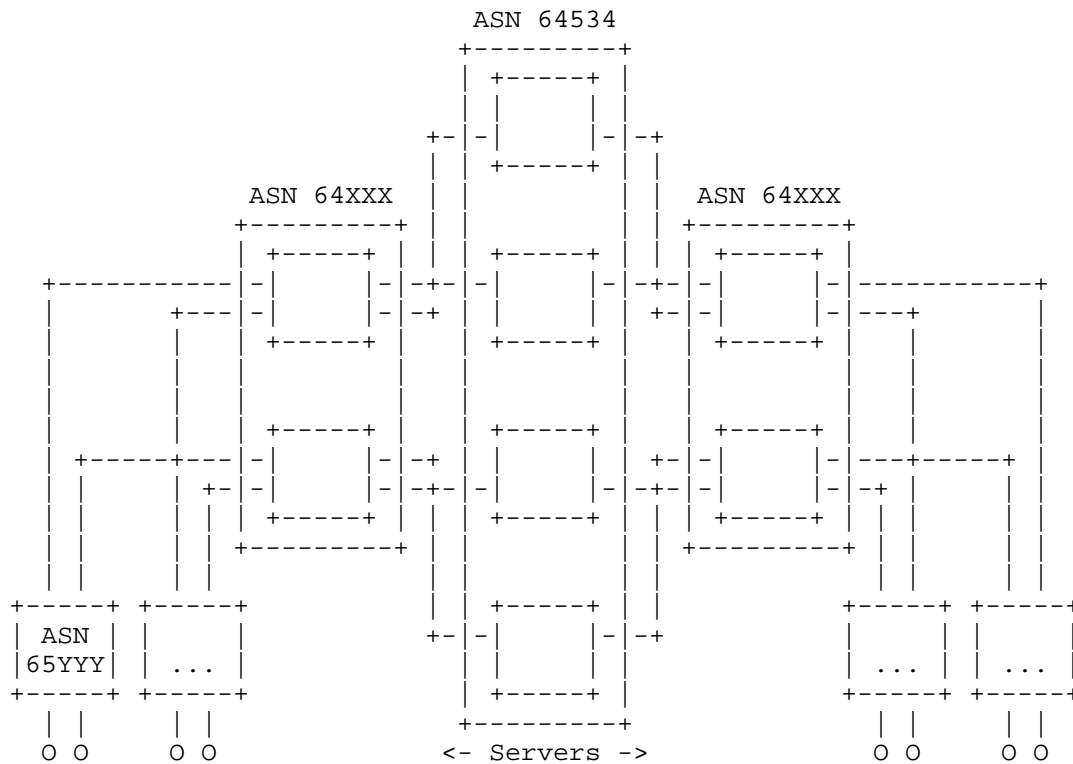


Figure 4: BGP ASN layout for 5-stage Clos

7.2.2. Non-unique private BGP ASN's

The use of private BGP ASNs limits to the usable range of 1022 unique numbers. Since it is very likely that the number of network switches could exceed this number, a workaround is required. One approach would be to re-use the private ASN's assigned to the Tier 3 switches across different clusters. For example, private BGP ASN's 65001, 65002 ... 65032 could be used within every individual cluster to be assigned to Tier 3 switches.

To avoid route suppression due to AS PATH loop prevention, upstream eBGP sessions on Tier 3 switches must be configured with the "AllowAS In" feature that allows accepting a device's own ASN in received route advertisements. Introducing this feature does not create the opportunity for routing loops under misconfiguration since the AS PATH is always incremented when routes are propagated from tier to tier.

Another solution to this problem would be to using four-octet (32-bit) BGP ASNs. However, there are no reserved private ASN range in the four-octet numbering scheme although efforts are underway to support this, see [I-D.mitchell-idr-as-private-reservation]. This will also require vendors to implement specific policy features, such as four-octet private AS removal from AS-PATH attribute.

7.2.3. Prefix advertisement

A Clos topology has a large number of point-to-point links and associated prefixes. Advertising all of these routes into BGP may create FIB overload conditions. There are two possible solutions that can help prevent FIB overload:

- o Do not advertise any of the point-to-point links into BGP. Since eBGP peering changes the next-hop address anyways at every node, distant networks will automatically be reachable via the advertising eBGP peer
- o Advertising point-to-point links, but summarizing them on every advertising device. This requires proper address allocation, for example allocating a consecutive block of IP addresses per Tier 1 and Tier 2 device to be used for point-to-point interface addressing.

Server facing subnets on Tier 3 switches are announced into BGP without using summarization on Tier 2 and Tier 1 switches. Summarizing subnets in the Clos topology will result in route black-holing under a single link failure (e.g. between Tier 2 and Tier 3 switch) and hence must be avoided. The use of peer links within the same tier to resolve the black-holing problem is undesirable due to $O(N^2)$ complexity of the peering mesh and waste of ports on the switches.

7.2.4. External connectivity

A dedicate cluster (or clusters) in the Clos topology could be used solely for the purpose of connecting to the Wide Area Network (WAN) edge devices, or WAN Routers. Tier 3 switches in such a cluster would be replaced with WAN Routers, but eBGP peering would be used again, though WAN routers are likely to belong to a public ASN.

The Tier 2 devices in such a dedicated cluster will be referred to as "Border Routers" in this document. These devices have to perform a few special functions:

- o Hide network topology information when advertising paths to WAN routers, i.e. remove private BGP ASNs from the AS-PATH attribute. This is typically done to avoid BGP ASN number collisions across

the data centers. A BGP policy feature called "Remove Private AS" is commonly used to accomplish this. This feature strips a contiguous sequence of private ASNs found in AS PATH attribute prior to advertising the path to a neighbor. This assumes that all BGP ASN's used for intra data center numbering are from the private ASN range.

- o Originate a default route to the data center devices. This is the only place where default route could be originated, as route summarization is highly undesirable for the "scale-out" topology. Alternatively, Border Routers may simply relay the default route learned from WAN routers.

7.3. ECMP Considerations

This section covers the Equal Cost Multipath (ECMP) functionality for Clos topology and discusses a few special requirements.

7.3.1. Basic ECMP

ECMP is the fundamental load-sharing mechanism used by a Clos topology. Effectively, every lower-tier switch will use all of its directly attached upper-tier devices to load-share traffic destined to the same prefix. Number of ECMP paths between two input/output switches in Clos topology equals to the number of the switches in the middle stage (Tier 1). For example, Figure 5 illustrates the topology where Tier 3 device A has four paths to reach servers X and Y, via Tier 2 devices B and C and then Tier 1 devices 1, 2, 3, and 4 respectively.

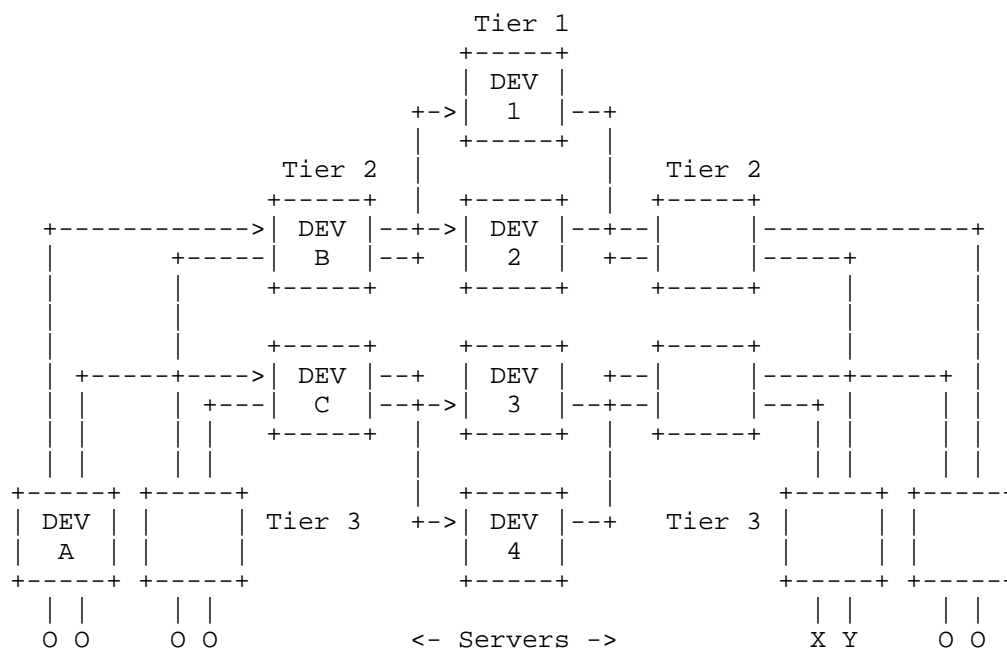


Figure 5: ECMP fan-out tree from A to X and Y

The ECMP requirement implies that the BGP implementation must support multi-path fan-out for up to the maximum number of devices directly attached at any point in the topology. Normally, this number does not exceed half of the ports found on a switch in the topology. For example, an ECMP max-path of 32 would be required when building a Clos network using 64-port devices.

Most implementations declare paths to be equal from ECMP perspective if they match up to and including step (e) in Section 9.1.2.2 of [RFC4271]. In the proposed network design there is no underlying IGP, so all IGP costs are automatically assumed to be zero (or otherwise the same value across all paths). Loop prevention is assumed to be handled by the BGP best-path selection process.

7.3.2. BGP ECMP over multiple ASN

For application load-balancing purposes we may want the same prefix to be advertised from multiple Tier-3 switches. From the perspective of other devices, such a prefix would have BGP paths with different AS PATH attribute values, though having the same AS PATH attribute lengths. Therefore, the BGP implementations must support load-sharing over above-mentioned paths. This feature is sometimes known

as "AS PATH multipath relax" and effectively allows for ECMP to be done across different neighboring ASNs.

7.4. BGP convergence properties

This section reviews routing convergence properties of BGP in the proposed design. A case is made that sub-second convergence is achievable provided that implementation supports fast BGP peering session shutdown upon failure of an associated link.

7.4.1. Convergence timing

BGP typically relies on an IGP to route around link/node failures inside an AS, and implements either a polling based or an event-driven mechanism to obtain updates on IGP state changes. The proposed routing design omits the use of an IGP, so the only mechanisms that could be used for fault detection are BGP keep-alives and link-failure triggers.

Relying solely on BGP keep-alive packets may result in high convergence delays, in the order of multiple seconds (normally, the minimum recommended BGP hold time value is 3 seconds). However, many BGP implementations can shut down local eBGP peering sessions in response to the "link down" event for the outgoing interface used for BGP peering. This feature is sometimes called as "fast fail-over". Since the majority of the links in modern data centers are point to point fiber connections, a physical interface failure is often detected in milliseconds and subsequently triggers a BGP re-convergence.

Furthermore, popular link technologies, such as 10Gbps Ethernet, may support a simple form of OAM for failure signaling such as [FAULTSIG10GE], which makes failure detection more robust. Alternatively, as opposed to relying on physical layer for fault signaling, some platforms may support Bidirectional Forwarding Detection ([RFC5880]) to allow for sub-second failure detection and fault signaling to the BGP process. This, however, presents additional requirements to vendor software and possibly hardware, and may contradict REQ1.

7.4.2. Failure impact scope

BGP is inherently a distance-vector protocol, and as such some of failures could be masked if the local node can immediately find a backup path. The worst case is that all devices in data center topology would have to either withdraw a prefix completely, or recalculate the ECMP paths in the FIB. Reducing the fault domain using summarization is not possible with the proposed design, since

using this technique may create route black-holing issues as mentioned previously. Thus, the control-plane failure impact scope is the network as a whole. It is worth pointing that such property is not a result of choosing BGP, but rather a result of using the "scale-out" Clos topology.

7.4.3. Third-party route injection

BGP allows for a third-party BGP speaker (not necessarily directly attached to the network devices) to inject routes anywhere in the network topology. This could be achieved by peering an external speaker using an eBGP multi-hop session with some or even all devices in the topology. Furthermore, BGP diverse path distribution [I-D.ietf-grow-diverse-bgp-path-dist] could be used to inject multiple next-hop for the same prefix to facilitate load-balancing. Using such a technique would make it possible to implement unequal-cost load-balancing across multiple clusters in the data-center, by associating the same prefix with next-hops mapped to different clusters.

For example, a third-party BGP speaker may peer with Tier 3 and Tier 1 switches, injecting the same prefix, but using a special set of BGP next-hops for Tier 1 devices. Those next-hops are assumed to resolve recursively via BGP, and could be, for example, IP addresses on Tier 3 switches. The resulting forwarding table programming could provide desired traffic proportion distribution among different clusters.

8. Security Considerations

The design does not introduce any additional security concerns. For control plane security, BGP peering sessions could be authenticated using TCP MD5 signature extension header [RFC2385]. Furthermore, BGP TTL security [I-D.gill-btsh] could be used to reduce the risk of session spoofing and TCP SYN flooding attacks against the control plane.

9. IANA Considerations

There are no considerations associated with IANA for this document.

10. Acknowledgements

This publication summarizes work of many people who participated in developing, testing and deploying the proposed design. Their names, in alphabetical order, are George Chen, Parantap Lahiri, Dave Maltz,

Edet Nkposong, Robert Toomey, and Lihua Yuan. Authors would also like to thank Jon Mitchell, Linda Dunbar and Susan Hares for reviewing and providing valuable feedback on the document.

11. Informative References

- [RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, December 2006.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [I-D.ietf-grow-diverse-bgp-path-dist]
Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of diverse BGP paths.",
draft-ietf-grow-diverse-bgp-path-dist-07 (work in progress), May 2012.
- [I-D.mitchell-idr-as-private-reservation]
Mitchell, J., "Autonomous System (AS) Reservation for Private Use", draft-mitchell-idr-as-private-reservation-00 (work in progress), June 2012.
- [I-D.gill-btsh]
Gill, V., Heasley, J., and D. Meyer, "The BGP TTL Security Hack (BTSH)", draft-gill-btsh-02 (work in progress), May 2003.
- [GREENBERG2009]
Greenberg, A., Hamilton, J., and D. Maltz, "The Cost of a Cloud: Research Problems in Data Center Networks", January 2009.
- [FAULTSIG10GE]
Frazier, H. and S. Muller, "Remote Fault & Break Link Proposal for 10-Gigabit Ethernet", September 2000.
- [INTERCON]
Dally, W. and B. Towles, "Principles and Practices of Interconnection Networks", ISBN 978-0122007514, January 2004.

[ALFARES2008]

Al-Fares, M., Loukissas, A., and A. Vahdat, "A Scalable,
Commodity Data Center Network Architecture", August 2008.

Authors' Addresses

Petr Lapukhov
Microsoft Corp.
One Microsoft Way
Redmond, WA 98052
US

Phone: +1 425 7032723 X 32723
Email: petrlapu@microsoft.com
URI: <http://microsoft.com/>

Ariff Premji
Arista Networks
5470 Great America Parkway
Santa Clara, CA 95054
US

Phone: +1 408-547-5699
Email: ariff@aristanetworks.com
URI: <http://aristanetworks.com/>

Network Working Group
Internet-Draft
Intended status: Informational
Expires: October 5, 2016

R. White
Linkedin
A. Retana
Cisco Systems, Inc.
S. Hares
Huawei
April 4, 2016

Filtering of Overlapping Routes
draft-white-grow-overlapping-routes-04

Abstract

This document proposes an optional mechanism to remove a prefix when it overlaps with a functionally equivalent shorter prefix. The proposed mechanism does not require any changes to the BGP protocol.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on October 5, 2016.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Requirements Language	3
3. Overlapping Route Filtering Mechanism	3
3.1. Marking Overlapping Routes	4
3.2. Preferring Marked Routes	4
3.2.1. Using a Cost Community	4
3.2.2. Using the Local Preference	4
3.3. Handling Marked Routes Within the AS	5
3.4. Handling Marked Routes at the Outbound Edge	5
4. Examples of Filtering Overlapping Routes	5
4.1. IPv4 Example	5
4.2. IPv6 Example	6
5. Operational Considerations	6
5.1. Advantages to the Service Provider	7
5.2. Implications for Router processing	7
5.3. Implications for Convergence Time	7
6. Security Considerations	7
7. IANA Considerations	8
8. Acknowledgements	8
9. References	8
9.1. Normative References	8
9.2. Informative References	8
Appendix A. Change Log	8
A.1. Changes between the -00 and -01 versions.	8
A.2. Changes between the -01 and -02 versions	9
A.3. Changes between the -02 and -03 versions	9
Authors' Addresses	9

1. Introduction

One cause of the growth of the global Internet's default free zone table size is overlapping routes injected into the routing system to steer traffic among various entry points into a network. Because padding AS Path lengths can only steer inbound traffic in a very small set of cases, and other mechanisms used to steer traffic to a particular inbound point are ineffective when multiple upstream providers are in use, advertising longer prefixes is often the only possible way for an AS to steer traffic into specific entry points along its edge.

These longer prefix routes, called overlapping routes in this document, are often advertised along with a shorter prefix route, called a covering route, in order to ensure connectivity in the case

of link or device failures. Overlapping routes not only add to the load on routers in the Internet core by simply expanding the table size; these routes may be less stable than the covering routes they are paired with.

Given the importance of an autonomous system's ability to steer traffic into specific entry points, simply removing the longer prefixes in a longer prefix (overlapping)/shorter prefix (covering) pair of routes isn't a viable solution.

This document proposes an optional mechanism to remove overlapping routes that are no longer useful for steering traffic towards a specific entry point in a particular AS. Removing these routes would reduce the global table in size, and reduce its instability, while removing no capabilities, nor increasing the average path length.

The mechanism proposed is simple to implement, requiring no changes to BGP [RFC4271] either in packet format or in the decision process. The removal described in this document is akin to filtering, not to route aggregation.

The intent of the mechanism is for it to be used based on local decisions and policies, not on an Internet-wide fashion. It is assumed that network operators using this mechanism have an incentive to do so.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Overlapping Route Filtering Mechanism

The handling of overlapping prefixes received from an external peer can be broken down into four parts: marking overlapping routes, preferring marked routes, handling marked routes within the AS, and handling marked routes at the AS exit point.

The initial step in successfully filtering overlapping routes is to identify and mark them. This document proposes the use of a BGP community called BOUNDED for that purpose. Because the operation suggested takes place inside an Autonomous System (AS), then any locally assigned community can be used.

The term BOUNDED is used to refer to a locally assigned community used to mark overlapping routes, and to these marked routes as well.

3.1. Marking Overlapping Routes

As each prefix is received by a BGP speaker from an external peer, it is evaluated in the light of other prefixes already received. If two prefixes overlap in space (such as 192.0.2.0/24 and 192.0.2.128/25, or 2001:DB8::/32 and 2001:DB8:1:/48), the longer prefix SHOULD be BOUNDED if it fully overlaps the covering prefix and it is the best path to the destination.

An overlapping prefix is said to fully overlap the corresponding covering prefix if both have identical AS_PATH attributes (both in length and contents) and the same NEXT_HOP.

3.2. Preferring Marked Routes

Since the same overlapping route may be received at several peering points along the edge of the AS, and the covering route may not be present at each of these points, BOUNDED routes SHOULD be preferred over unmarked routes for overlapping routes to be properly handled. A router which marks an overlapping route should also use one of the two mechanisms described here to insure the marked route is preferred throughout the AS.

Only one method described in this section SHOULD be deployed in any given AS.

3.2.1. Using a Cost Community

The recommended method for preferring BOUNDED routes is to use a Cost Community [I-D.ietf-idr-custom-decision] with the Point of Insertion set to ABSOLUTE_VALUE. This mechanism leaves all existing local policy controls in place within the AS.

If this method is used, only the BOUNDED routes need to be tagged using a lower than default Cost, as routes without a Cost Community are considered to have the default value.

3.2.2. Using the Local Preference

An alternate mechanism which may be used to prefer BOUNDED routes is to set their Local Preference to some number higher than the normal standard policy settings for a particular prefix. It's not important that any particular BOUNDED route win over any other one; so simply adding a small amount to the normal Local Preference, as dictated by local policy, will ensure a BOUNDED route will always win over an unmarked route, so only these routes reach the outbound edge of the AS.

3.3. Handling Marked Routes Within the AS

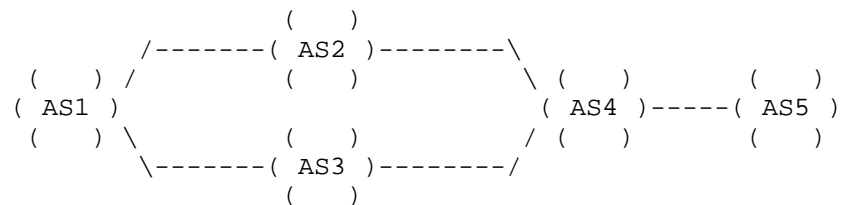
Routes marked with the BOUNDED community MAY not be installed in the local RIB of routers within the AS. This optional step will reduce local RIB and forwarding table usage and volatility within the AS.

3.4. Handling Marked Routes at the Outbound Edge

If local policy dictates, routes marked with the BOUNDED community SHOULD NOT be advertised to external peers. If they are advertised, they MAY then be marked with the NO_EXPORT community.

4. Examples of Filtering Overlapping Routes

Assume the following configuration of autonomous systems:



This network is used in both of the following examples.

4.1. IPv4 Example

- o AS1 is advertising 192.0.2.128/25 to both AS2 and AS3.
- o AS2 is advertising both 192.0.2.128/25 and 192.0.2.0/24 into AS4.
- o AS3 is advertising 192.0.2.128/25 into AS4
- o Each BGP connection (session) is handled by a separate router within each AS (for instance, AS4 peers with AS2 and AS3 on separate routers).

When the router in AS4 peering with AS2 receives both the 192.0.2.128/25 and the 192.0.2.0/24 prefixes, it will mark 192.0.2.128/25 as BOUNDED, and set a Cost Community (as described in Section 3.2.1) so the marked overlapping route is preferred over unmarked routes within AS4.

The border router between AS4 and AS3 will receive the longer prefix from AS3, and the preferred BOUNDED overlapping route through iBGP. It will prefer the marked route, so the unmarked route towards 192.0.2.128/25 will not be advertised throughout AS4.

If the link between AS1 and AS2 fails, the longer length prefix will be withdrawn from AS2, and thus the peering point between AS2 and AS4 will no longer have an overlapping set of prefixes. Within AS4, the border router which peers with AS2 will cease advertising the 192.0.2.128/25 prefix, which allows the AS3/AS4 border router to begin advertising it into AS4, and through AS4 into AS5, restoring connectivity to AS1.

4.2. IPv6 Example

- o AS1 is advertising 2001:DB8:1:/48 to both AS2 and AS3.
- o AS2 is advertising both 2001:DB8:1:/48 and 2001:DB8::/32 into AS4.
- o AS3 is advertising 2001:DB8:1:/48 into AS4
- o Each BGP connection (session) is handled by a separate router within each AS (for instance, AS4 peers with AS2 and AS3 on separate routers).

When the router in AS4 peering with AS2 receives both the 2001:DB8:1:/48 and 2001:DB8::/32 prefixes, it will mark 2001:DB8:1:/48 as BOUNDED, and set a Cost Community (as described in Section 3.2.1) so the marked overlapping route is preferred over unmarked routes within AS4.

The border router between AS4 and AS3 will receive the longer prefix from AS3, and the preferred BOUNDED overlapping route through iBGP. It will prefer the marked route, so the unmarked route towards 2001:DB8:1:/48 will not be advertised throughout AS4.

If the link between AS1 and AS2 fails, the longer length prefix will be withdrawn from AS2, and thus the peering point between AS2 and AS4 will no longer have an overlapping set of prefixes. Within AS4, the border router which peers with AS2 will cease advertising the 2001:DB8:1:/48 prefix, which allows the AS3/AS4 border router to begin advertising it into AS4, and through AS4 into AS5, restoring connectivity to AS1.

5. Operational Considerations

The intent of the mechanism described in this document is for it to be used based on local policies, not on an Internet-wide fashion. It is assumed that network operators using this mechanism have an incentive to do so.

The practice of filtering exists today on the Internet. While there may be local benefits to applying manual filters and/or the mechanism

specified in this document, the operator should be aware of the impact it may have on neighboring autonomous systems' policies [I-D.cardona-filtering-threats].

The benefits and implications associated with this proposal are discussed in the sections below. The text references the sample network in Section 4.

5.1. Advantages to the Service Provider

AS4, in each of the situations, reduces the number of prefixes advertised to transit peering autonomous systems by the number of longer prefixes that overlap with aggregates of those prefixes, so that AS5 receives fewer total routes, and a more stable routing table. While one copy of the prefix continues to be carried through the autonomous system, this entry can be removed from the local forwarding table.

5.2. Implications for Router processing

This proposal requires a BGP speaker to perform an additional check on receiving a route, checking the route against existing routes for overlapping coverage of a set of reachable destinations. This additional work, in terms of processing requirements, should be easily offset by the overall savings in processing through the reduction of the forwarding table size, and the additional stability in the routing table due to the removal of longer length prefixes.

5.3. Implications for Convergence Time

If the route to the AS providing the route to the covering route should be lost, the overlapping route must now propagate into the autonomous systems which had formerly received only the covering route. This behavior increases convergence time and may create situations in which reachability is temporarily compromised. Unlike the case where manual filters are used, normal BGP behavior should restore reachability without changes to the router configuration.

6. Security Considerations

This document presents a mechanism for an autonomous system to mark and filter overlapping prefixes. Note that the result of this operation is akin to the implementation of local route filtering at an AS boundary. As such, this document doesn't introduce any new security risks.

7. IANA Considerations

This document has no IANA actions.

8. Acknowledgements

Cengiz Alaentnoglu, Daniel Walton, David Ball, Ted Hardie, Jeff Hass, Barry Greene, Bill Herrin and Robert Raszuk gave valuable comments on this document.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2. Informative References

[I-D.cardona-filtering-threats]
Cardona, C. and P. Francois, "Making BGP filtering a habit: Impact on policies", draft-cardona-filtering-threats-02 (work in progress), July 2013.

[I-D.ietf-idr-custom-decision]
Retana, A. and R. White, "BGP Custom Decision Process", draft-ietf-idr-custom-decision-04 (work in progress), November 2013.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

Appendix A. Change Log

A.1. Changes between the -00 and -01 versions.

- o Updated authors' contact information.
- o Changed intended status to Informational.
- o General editorial changes.
- o Clarified the intent of the draft in several places.
- o Clarified when a route should be marked (3.1).
- o Edited the operational considerations section.

- o Updated ACKs.
- A.2. Changes between the -01 and -02 versions
 - o Updated authors' contact information.
 - o General editorial changes.
 - o Refined the text about marking routes.
- A.3. Changes between the -02 and -03 versions
 - o Updated authors' contact information.
 - o Added IPv6 examples.
 - o Minor editorial changes.
- A.4. Changes between the -03 and -04 versions
 - o Updated authors' contact information.

Authors' Addresses

Russ White
Linkedin

Email: russ@riw.us

Alvaro Retana
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
USA

Email: aretana@cisco.com

Susan Hares
Huawei

Email: shares@ndzh.com